Learning speaker-specific pronunciations of disordered speech

H. Christensen, P. Green, T. Hain

Computer Science, University of Sheffield, Sheffield, United Kingdom

{h.christensen, p.green, t.hain}@dcs.shef.ac.uk

Abstract

One of the main clinical applications of speech technology is in voice-enabled assistive technology for people with disordered speech. Progress in this area is hampered by a sparseness in suitable data and recent research have focused on ways of incorporating knowledge about typical (i.e., un-impaired) speech through the use of e.g., deep belief neural networks. This paper presents a new way of using deep belief neural networks trained on typical speech, namely to improve pronunciations for individual speakers. Analysis of the posterior probabilities show a clear correlation between measured pronunciation 'disorderedness' and the overall speech recognition performance of the full system. Based on this, we propose a method to use deep belief network outputs to i) identify which words are pronounced differently than what would be expected from a typical pronunciation, and ii) subsequently generate new pronunciations. We investigate different methods for pronunciation generation as well as what is the best way of using the modified pronunciations to inform the system development stages. Using the UAspeech database of disordered speech, we demonstrate improvement in average accuracy of 69.76% to 70.51%, with some speakers showing individual improvements of up to 10%.

Index Terms: dysarthric speech recognition, learning speaker-specific pronunciations

1. Introduction

People with severe physical handicaps may find it difficult to operate devices in their homes. Providing hands-free, voice-enabled assistive technology can help towards them experiencing more independence and potentially lead to a better quality of life [1]. However, as many physical handicaps affect motor-control mechanisms including the speech articulators, potential users often have *disordered* speech, also known as *dysarthric* speech. Dysarthric speech is characterised by being 'slurred' as well as highly variable, and can be hard to understand for listeners unfamiliar with the particular impairments of a speaker. This often rules out the use of conventional, off-the-shelf automatic speech recognition (ASR) systems trained on *typical* speech [2], leaving clinical scientists and researchers working with dysarthric speakers with many design choices in order to establish a system that is tailored to the specifics of a given speaker's impairment, yet leverages off typical speech ASR knowledge as much as possible.

One of the things that impacts on choice of models and training strategies is the lack of suitable training data as speakers with dysarthria can find it tiring and distressing to speak for any length of time. Very few databases exists, and research often concentrates on optimising learning strategies when faced with sparse data. Some success have been demonstrated by using both whole-word and triphone models in speaker dependent and speaker adapted systems [3, 4, 1, 5, 6]. In [7] we trained dysarthric triphone-based speech ASR systems by applying 'state-of-the-art' training and adaptation strategies developed on typical speech, and achieved significant improvements on previously published results using the UAspeech database (one of the largest databases of English dysarthric speech [8]). In [9] we have further expanded on this work by including features generated from deep belief neural networks (DNNs) [10] trained on *out-of-domain* (OOD) data, such as typical speech data from the AMI collection of meeting data. Adding features from the DNNs pre-trained on OOD provides a further improvement for the UAspeech task. However, for some speakers using conventional training strategies and typical speech data is only going to help up to a certain point: their pronunciation patterns are too different from that which is inherently expected of a system with a canonical dictionary. This paper addresses this issue by looking at ways of tailoring an ASR system to a speaker. We propose a method of altering the training data transcriptions and test data dictionaries so they are tuned to the individual speaker and matches the particular speech impairment of that speaker.

In a conventional ASR system, the word pronunciations are normally fixed and unlike the acoustic and language models, often no attempt is made to derive pronunciations using data-driven methods. Exceptions are automatic methods for generating pronunciations for out-of-vocabulary words, letter-to-sound mapping [11, 12, 13], and learning accent and dialect specific dictionaries [14, 15]. In the area of disordered speech, research on the detection of pronunciation errors have attracted some interest [16, 17], as have work on predicting intelligibility [2, 18, 19]. Mengistu and Rudzicz showed good results by adapting the pronunciation dictionaries to individual speakers [20]. However, their method involve expert assessment of the individual speakers' pronunciation pattern which can be very time-consuming.

This paper presents an automatic way of detecting mispronunciations and subsequently adjusting the dictionary to reflect this. We look at using the posterior probability outputs of OOD DNNs as indicators of how a particular speaker's pronunciation of a given word is different from that represented in the typical speech pronunciation. We investigate the different ways of letting this information feed back into the training, adaptation and testing stages of a system.

In section 2 we quantify the degree of discrepancy between the dysarthric speech pronunciations as observed in the posterior probabilities in comparison with those of a typical speaker, and we analyse to what degree this discrepancy is correlated with the overall ASR performance for a given speaker. Section 4 gives some details about the experimental setup as well as details of the different specific strategies we have explored for deriving alternative speaker dependent pronunciations, and finally sections 5 and 6 present our results and conclusions.

2. Analysis

We base our analysis on DNNs pre-trained on typical speech from the AMI corpus. They are trained using mono-phone label targets and by applying a suitable sigmoid function at the final layer we can interpret the output as frame-level posterior probabilities. When passing dysarthric speech from a particular speaker through the network it is hence possible to interprete peaks in the posterior probabilities as indicators of the pronunciation pattern. We investigate to which degree the DNNs can be used to identify pronunciation mismatch from that of typical speech, and how better pronunciations can be derived.

Figure 1, top frame, shows a typical example of the posterior probabilities output from a DNN; this is for all of the frames from the word 'Line' spoken by speaker F04 from the UAspeech database. The dark areas indicate phone labels that have a very high probability. Moving along the x-axis, it is possible to see a relatively clear sequence of likely phone labels, with some uncertainty around the phone label '/l/' where it looks like the speaker has sounded intermittently like an '/m/' or '/n/'. For each frame, finding the phone label with the maximum posterior probability would yield the most likely phone sequence, with the highest possible probability for each frame. One can find the frame-level posterior probability sequence for any phone sequence by indexing into the posterior probability matrix. The resulting vector of posterior probabilities will have values close to 1 in frames where the phone label with which we are indexing is a good match to how the phone is pronounced, and likewise have low probabilities if there is a big mismatch.

The bottom pane of Figure 1 shows the posterior probabilities for two such different phone sequences, namely, i) the sequence corresponding to the pronunciations from the standard, speaker independent dictionary (based on typical speech pronunciation patternsand ii) the phone sequence which is the most likely when looking at which phone label has the maximum probability per frame; the maximum sequence can be seen as an 'observed' (or hypothesised) pronunciation for this particular speaker and this particular word instance.

The speaker in Figure 1 has pronounced all of the phonemes in the word fairly correctly except /l/. This is reflected in the line plot of the posterior probabilities for the two phone sequences: the probability for the typical speech pronunciation drops around frames 20-25, and there is a *discrepancy* between the posterior probabilities of the typical speech pronunciation and those corresponding to the observed/dysarthric speech pronunciation. We hypothesise that the size of the distance between the observed and typical phone sequence is an indicator of to what degree the speaker's speech is disordered. To investigate this further, we accumulated the distance between observed and typical probabilities over all words said by each speaker. We then looked at the correlation between this value and the intelligibility of the individual speakers as well as the WER achieved for the full ASR systems.

Figure 2 illustrates this. It contains a plot of the speaker dependent measured discrepancies vs. either WER for the full ASR system (crosses on the left-hand plot) or the speaker's intelligibility as provided with the UAspeech meta-data (crosses on the right-hand plot). There is a clear correlation between probability mass distribution discrepancies and both the WER and the intelligibility. On each plot, a fitted linear regression curve is also plotted. The parameters corresponding to this line in the WER-plot accounts for 90% of the variance. The equivalent number for the intelligibility line is 87%. In summary, there is a strong correlation between the discrepancies as observed for

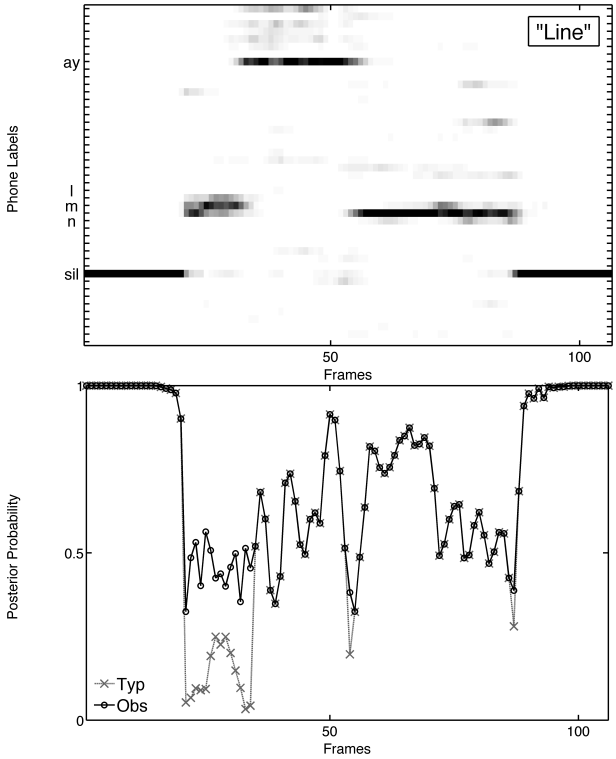

Figure 1: *Phone-level output posterior probabilities for all frames in the word 'Line'. For clarity, only phone labels relevant to this word is shown on the y-axis. The bottom plot shows two sequences of posterior probabilities for the same word. Please see text for further explanation.*

the posterior probability outputs and the resulting overall system WER. For a particular speaker in need of an ASR system, it is therefore possible to use a DNN in a diagnostic fashion to measure the 'disorderedness' of a particular speaker. Notably, it is possible that even very small amounts of data could give useful insight and be used as a first step towards tailoring that speaker's system configuration.

3. Method

We propose a method for deriving phone level pronunciations for individual speakers with dysarthric speech, which is based on using a DNN to find a phone sequence that has a higher accumulated posterior probability associated with it, than the typical phone sequence does.

Let $p_w(f, l)$ be the phone level posterior probability for word w, frame f and phone label l, and let L represent phone sequences arising from using a particular method of pronunciation generation, the accumulated posterior probability associated with that phone sequence L is given as

$$K_L = \sum_{w=1}^{W} \sum_{f=1}^{F_w} p_w(f, l) \Big|_{l \in L} \quad (1)$$

where W is the number of words for that speaker and F_w is the number of frames for word w. The difference in accumulated posterior probability for two pronunciation generation methods (e.g., pronunciations derived according to the typical

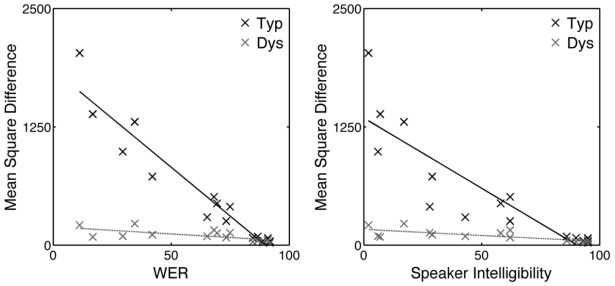

Figure 2: *Illustration of correlation between the degree of discrepancy between 'typ' and 'dys' (observed on dysarthric speech) posterior probability masses and the overall full system WER (lefthand-side panel) and the intelligibility for that speaker (righthand-side panel). Super-imposed on the plots are the linear regression lines.*

speech dictionary ('typ') and by doing a frame-level maximum the observed pronunciations described above ('dys')) is then:

$$\Delta K = K_{L_{dys}} - K_{L_{typ}} \tag{2}$$

$K_{L_{dys}}$ represent the phone sequence giving the maximum posterior probability, and we hypothesise that by choosing a pronunciation generation method, $\hat{L}$ which reduces the difference in Eq. 2, we can increase the performance of the final ASR system. This can be shown to be equivalent to reducing the distance for each word instance:

$$\begin{aligned} \min_{\hat{L}} \Delta K &= \min_{\hat{L}} \left\{ K_{L_{dys}} - K_{\hat{L}} \right\} \\ &= \min_{\hat{L}} \left\{ \sum_{w} \left(K_{L_{dys_w}} - K_{\hat{L}_w} \right) \right\} \\ &= \sum_{w} \min_{\hat{L}} \left(K_{L_{dys_w}} - K_{\hat{L}_w} \right) \end{aligned} \tag{3}$$

4. Experimental setup

We are looking at finding pronunciation generation methods, $\hat{L}$ which will improve recognition results for dysarthric speakers. Each tested method will be used to derive pronunciations from the training data, and the resulting transcriptions and dictionary will be applied at some or all of training, adaptation and testing stages of a full ASR system to investigate what the effect is.

To this end, we train a range of speaker independent and speaker adapted systems, which allows us to compare results to a set of baseline systems established by using the typical, speaker-independent annotations and pronunciation forms. We test the systems on the UAspeech task which is single word recognition and have continued following the decoding strategy deployed in [7, 9]. A uniform language model was used, with a word grammar network containing silence models at the start and end, and all possible test words in parallel.

For the acoustic modelling in all the systems, the data is encoded in 'tandem features'. These are feature vectors comprised of a conventional feature vector - in our case a 13-dimensional PLP vector with added first and second order derivatives - augmented with features extracted from a DNN pre-trained on the AMI corpus [21]. The DNN was trained on filterbank outputs and the DNN features used in the ASR systems are stacked bottleneck features as described in [22]. This system setup gave very competitive results in [9]. All Hidden Markov Models (HMMs) were trained using the maximum likelihood (ML) criterion. State-clustered, triphones having Gaussian mixture models with 16 components per state were used. Speaker adaptation was carried out using the maximum a posteriori (MAP) technique [23].

4.1. Data

The UAspeech database contains synchronised audio and visual streams from 15 speakers (4 female and 11 male). The dysarthric speakers were asked to repeat single words from 5 groups: 10 digits, 29 Nato alphabet letters, 19 command words ('delete', 'enter' etc.), 100 common words ('the','will' etc.), and 300 uncommon words chosen to be phonetically rich and complementary to the remaining words ('Copenhagen','chambermaid' etc.). Full details of the corpora can be found in [8]. The speakers all have dysarthric speech, and accompanying the database are percent intelligibly scores as obtained from listening tests with unfamiliar listeners. These range from 4% to 95%. Following previously published work using the UAspeech for ASR (e.g. [3]) the data was divided into training and test data with a 2:1 split, using blocks 1 and 3 for training and block 2 for testing. As our method relies on having examples of a particular word in the training set, we exclude all the uncommon words which differ between each recording block. In total we have about 9 and 3 hours per speaker in the training and test set respectively.

The DNN was trained on the AMI dataset [22], which consists of meeting room headset microphone recordings with multiple speakers per session. It contains about 127 hours of data. The speech is conversational of nature and there is a relatively large variety in accent, (although all speakers can be considered fluent in English).

5. Results

We explored the following methods for generating speaker dependent pronunciations:

'raw' : In the extreme, setting $\hat{L} == L_{dys}$ in Eq. 2 would minimise the distance but also render the pronunciations practically useless by becoming very long. In practice, we generated the pronunciations and subsequently applied post-filtering by removing all phones that were too short (affected frames were then re-assigned to neighbouring phones). A limit of 4 frames was found empirically to be a good choice.

'avrmax' : One way of reducing the number of phones in each pronunciation is to impose some structure. In all the following methods, this is done by imposing the phone boundaries from the reference alignments for that particular speaker and word, whilst still allowing the phone label identity to change.[1]. For each phone segment, n, the 'avrmax' method assigns the phone label, l_n as the one with the highest posterior probability, i.e. $l_n = \max_{l} = K_{L_{dys_n}}$. Best results were obtained by being conservative about which segments were given new phone labels, and the results presented here will keep the reference phone label *except* when the K-distance is too large (i.e., we are confident that the speakers pronunciation deviates from the typical). The threshold was empirically set at 0.8.

[1]Note, that this a step towards better pronunciations modelling: phoneme substitutions are handled, however further work is needed to expand the methods to also allow for deletion and insertions

	Word Accuracy	Pronunciations per word
'typ'	69.99	1.16
'raw'	59.96	6.38
'avrmax'	70.37	4.32
'probwords5'	70.51	1.17
'probwords20'	70.38	1.17
'typocc'	70.37	1.16

Table 1: *Word accuracy as a result of using different types of speaker specific dictionaries during testing and number of pronunciations per word. Please see text for explanation of pronunciation generation methods.*

'probwords' : This method extends the 'avrmax' methods above by restricting it to a subset of the words that appear to be *problematic*, as in mis-recognised when tested on a part of the training set. The size of the subset will greatly affect the performance of this method and we evaluated a large range of values. Here we present results from two such evaluations: either allowing new pronunciations for the top 5 or the top 20 problem words.

'typocc' : This method is the most conservative of the methods in that it will use – for each speaker – the actual pronunciation picked during a forced alignment for a particular word from the typical speech dictionary. The effect is that no new pronunciations are generated, however, pronunciation variations present in the typical speech dictionary which do *not* appear to be used by the speaker will not be present in the speaker dependent dictionary.

5.1. Effect of different pronunciation generation methods

The results of testing the various pronunciation generation methods is presented in Table 1 alongside the average number of pronunciations in the generated dictionaries. The 'typ' dictionary is the typical speech dictionary, i.e., the un-modified, speaker independent dictionary.

The 'raw' method is by far the worst with an accuracy of 59.96%, which is about 10% lover than that of the baseline system ('typ'). Although the 'raw' method produces dictionaries with pronunciations very close to the pronunciations observed, they clearly fail to generalise. The resulting pronunciations tend to be very long phone sequences with an average of 6.21 words which is almost twice as long as for the other methods that were all restricted to using the same number of phonemes as the 'typ' pronunciations had. All of the other methods show an improvement in performance in comparison to the 'typ' score, with the 'probwords5' giving the highest accuracy.

5.2. Effect of using speaker dependent pronunciations at training, adaptation and testing stages

We also wanted to investigate the effect of using the pronunciations during training and adaptation, and so we ran all possibly combinations of using typical, reference pronunciations ('typ') or speaker dependent, dysarthric speech pronunciations ('DYS'). Only the results for the 'probwords5' pronunciation generation methods are shown in Table 2, but we observed the same trends for all the methods.

Looking first at the overall averages, it can be seen that the earlier the 'DYS' pronunciations are introduced the better: the results gradually increase from the baseline result of 65.91% to 70.51%. Inspecting the results for the individual speakers show a more detailed picture: as expected some speakers have not improved (a few even have a negative result) whereas other speakers have improved by up to 10%. This confirms our initial statement that introducing speaker specific dictionaries can greatly benefit some speakers.

Train		typ	typ	typ	typ	DYS	DYS
Adapt		-	-	typ	typ	DYS	DYS
Test		typ	DYS	typ	DYS	typ	DYS
M04	(2%)	10.0	10.4	12.9	13.1	12.8	12.9
F02	(6%)	39.0	39.0	42.9	44.2	46.8	47.0
M12	(7%)	21.5	22.1	27.5	27.8	26.0	26.3
M01	(17%)	39.7	41.7	49.4	49.9	46.6	48.0
M07	(28%)	80.5	80.1	83.8	83.7	85.0	84.6
F02	(29%)	50.7	51.9	60.0	61.1	63.9	63.5
M16	(43%)	70.1	71.3	73.3	73.1	75.6	76.9
M05	(58%)	70.0	70.4	74.9	76.5	75.9	76.5
M11	(62%)	73.9	73.1	76.0	74.9	74.5	73.6
F04	(86%)	76.4	77.0	80.8	81.7	83.1	84.0
M09	(90%)	89.9	89.8	91.4	90.9	91.6	91.6
M14	(93%)	83.7	83.7	87.0	86.8	87.2	86.9
M10	(95%)	94.8	94.7	96.2	96.3	96.4	95.9
M08	(95%)	92.7	92.1	94.4	93.9	94.1	93.5
F05	(95%)	95.8	96.2	95.9	95.9	96.4	96.4
Average		65.91	66.23	69.76	69.99	70.39	70.51

Table 2: *Word accuracy rates using the 'probwords5' pronunciation generation method. Results per speaker (intelligibility in parenthesis). Top three rows indicate usage (during training, adaptation and/or testing): 'typ' - typical speech pronunciations; 'DYS' - dysarthric speaker specific pronunciations.*

6. Discussion and conclusions

This paper has investigated one way of designing an ASR system for a speaker with dysarthria, namely by modifying the pronunciations so they represent the specific speech impairments of the speaker. We used DNNs pre-trained on typical speech and demonstrated that there exists a correlation between the phone-level posterior probabilities corresponding to the typical speech pronunciation and the performance (as measured by WER) of the final system. We have also seen how this discrepancy depends on the severity of the speaker's dysarthria and believe this finding can be utilised in a diagnostic fashion by practitioners and researchers looking at finding an optimal system for a particular speaker with dysarthria.

We then investigated several methods for generating speaker specific pronunciations and observed that overall best results were achieved by carefully choosing for which phones alternative pronunciations should derived. We also showed that best results are achieved when using the pronunciations during both training, adaptation and testing.

7. Acknowledgements

This research was supported by the EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology). The authors wish to Pawel Swietojanski of University of Edinburgh for his advice and help generating the posterior probabilities.

8. References

[1] M. S. Hawley, P. Enderby, P. Green, S. Cunningham, S. Brownsell, J. Carmichael, M. Parker, A. Hatzis, P. O'Neill, and R. Palmer, "A speech-controlled environmental control system for people with severe dysarthria," *Medical Engineering & Physics*, vol. 5, no. 29, pp. 586 – 93, 2007.

[2] L. Ferrier, H. Shane, H. Ballard, T. Carpenter, and A. Benoit, "Dysarthric speakers intelligibility and speech characteristics in relation to computer speech recognition." *Augmentative and Alternative Communication*, vol. 11, pp. 165–174, 1995.

[3] H. V. Sharma and M. Hasegawa-Johnson, "State transition interpolation and map adaptation for hmm-based dysarthric speech recognition," in *HLT/NAACL Workshop on Speech and Language Processing for Assistive Technology (SLPAT)*, 2010, pp. 72–79.

[4] H. V. Sharma and M. Hasegawa-Johnson, "Acoustic model adaptation using in-domain background models for dysarthric speech recognition," *Computer Speech and Language*, 2012.

[5] H. Christensen, S. Siddharth, P. O'Neill, Z. Clarke, S. Judge, S. Cunningham, and M. Hawley, "SPECS - an embedded platform, speech-driven environmental control system evaluated in a virtuous circle framework," in *In proc. Workshop on Innovation and Applications in Speech Technology*, 2012.

[6] F. Rudzicz, "Comparing speaker-dependent and speaker-adaptive acoustic models for recognizing dysarthric speech," in *Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility*, 2007, pp. 255–256.

[7] H. Christensen, S. Cunningham, C. Fox, P. Green, and T. Hain, "A comparative study of adaptive, automatic recognition of disordered speech," in *Proc Interspeech 2012*, Portland, Oregon, US, Sep 2012.

[8] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *Proceedings of Interspeech*, Brisbane, Australia, 2008, pp. 22–26.

[9] H. Christensen, M. B. Aniol, P. Bell, P. Green, T. Hain, S. King, and P. Swietojanski, "Improving recognition of disordered speech with out-of-domain knowledge," in *Interspeech'13*, 2013.

[10] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, Nov 2012.

[11] I. McGraw, I. Badr, and J. Glass, "Learning lexicons from speech using a pronunciation mixture model," *Trans. ASLP*, vol. 21, no. 2, pp. 357–366, 2013.

[12] H. Strik and C. Cucchiarini, "Modeling pronunciation variation for asr: A survey of the literature," *Speech Communication*, vol. 29, pp. 225–246, 1999.

[13] H. Strik, "Pronunciation adaptation at the lexical level," in *Proceedings of ITRW on Adaptation Methods for Speech Recognition*, 2001.

[14] V. Pagel, K. Lenzo, and A. Black, "Letter to sound rules for accented lexicon compression," in *Proceeding ICSLP'98*, 1998.

[15] M. Elmahdy, M. Hasegawa-Johnson, and E. Mustafawi, "Hybrid phonemic and graphemic modeling for arabic speech recognition," *International Journal of Computational Linguistics*, vol. 3, no. 1, 2012.

[16] S. O. C. Morales and S. Cox, "Modelling errors in automatic speech recognition for dysarthric speakers," *EURASIP Journal on Advances in Signal Processing*, no. 2, Jan 2009.

[17] W. K. Seong, J. Park, and H. Kim, "Dysarthric speech recognition error correction using weighted finite state transducers based on contextdependent pronunciation variation," in *Computers Helping People with Special Needs*, ser. Lecture Notes in Computer Science, K. Miesenberger, A. Karshmer, P. Penaz, and W. Zagler, Eds. Springer Berlin Heidelberg, 2012, vol. 7383, pp. 475–482.

[18] C. Middag, J.-P. Martens, and G. V. N. M. D. Bodt, "Automated intelligibility assessment of pathological speech using phonological features," *EURASIP Journal on Advances in Signal Processing*, no. 2, Jan 2009.

[19] S.-C. Yin, R. C. Rose, O. Saz, and E. Lleida, "Verifying pronunciation accuracy from speakers with neuromuscular disorders," in *Proceedings of Interspeech'08*, 2008.

[20] K. T. Mengistu and F. Rudzicz, "Adapting acoustic and lexical models to dysarthric speech," in *Proceedings of ICASSP'11*, 2011, pp. 4924–4927.

[21] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proceedings of ICASSP'00*, Istanbul, Turkey, June 2000, pp. 1635–1630.

[22] T. Hain, L. Burget, J. Dines, P. N. Garner, F. Grezl, A. el Hannani, M. Huijbregts, M. Karafiat, M. Lincoln, and V. Wan, "Transcribing meetings with the amida systems," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 2, pp. 486–498, 2011.

[23] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.